arXiv:2508.05385v1 [cs.SD] 7 Aug 2025

A Scalable Pipeline for Enabling Non-Verbal Speech Generation and Understanding

Runchuan Ye[1]*, Yixuan Zhou[1], Renjie Yu[1], Zijian Lin[1], Kehan Li[1], Xiang Li[1], Xin Liu[2], Guoyang Zeng[2], Zhiyong Wu[1]

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] ModelBest Inc, Beijing, China

Abstract

Human spoken communication involves not only lexical content but also non-verbal vocalizations (NVs) such as laughter, sighs, and coughs, which convey emotions, intentions, and social signals. However, most existing speech systems focus solely on verbal content and lack the ability to understand and generate such non-verbal cues, reducing the emotional intelligence and communicative richness of spoken interfaces. In this work, we introduce **NonVerbalSpeech-38K**, a large and diverse dataset for non-verbal speech generation and understanding, collected from real-world media and annotated using an automatic pipeline. The dataset contains 38,718 samples (about 131 hours) with 10 categories of non-verbal cues, such as laughter, sniff, and throat clearing. We further validate the dataset by fine-tuning state-of-the-art models, including F5-TTS and Qwen2-Audio, demonstrating its effectiveness in non-verbal speech generation and understanding tasks. Our contributions are threefold: (1) We propose a practical pipeline for building natural and diverse non-verbal speech datasets; (2) We release a large-scale dataset to advance research on non-verbal speech generation and understanding; (3) We validate the dataset's effectiveness by demonstrating improvements in both non-verbal speech synthesis and captioning, thereby facilitating richer human-computer interaction.

Datasets and Demos —
https://github.com/nonverbalspeech38k/nonverspeech38k

1 Introduction

Spoken interaction has become a core interface in human-computer communication, driven by the rapid advancement of automatic speech recognition (ASR), text-to-speech (TTS) synthesis, and large language models (LLMs). These technologies have enabled systems to understand and generate semantically coherent spoken content with impressive fluency (Zhang et al. 2023; Xie and Wu 2024; Fang et al., 2024; Défossez et al. 2024). However, human communication is far richer than just words. **Non-verbal vocalizations (NVs)**, such as laughter, sighs and other acoustic events not tied to lexical meaning, are crucial for conveying emotion, intentions, and social signals (Truong and Van Leeuwen 2007; Cortes et al. 2021; Cowen et al. 2019; Kamiloğlu, Fischer, and Sauter 2020; Lima et al. 2014). Despite their importance, most current speech systems lack the ability to perceive or express such non-verbal elements, a limitation that hinders the creation of truly human-like and empathetic interactions.

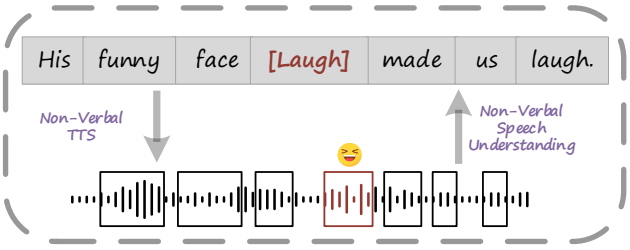

Figure 1: Task definition of non-verbal TTS and non-verbal speech understanding.

For an ideal speech system to genuinely mimic human communication, it must be capable of both understanding and generating non-verbal sounds. In this work, we focus on two key tasks that reflect these capabilities: non-verbal speech generation and non-verbal speech understanding (Figure 1).

Speech synthesis with NVs can be categorized into two distinct approaches. The first is non-verbal tag-controlled TTS, which inserts explicit tags (e.g., `[laughing]`, `[breath]`) into text to control the type and position of NVs (Guo et al. 2024; Du et al. 2024a,b, 2025; 2noise 2024; Nari-labs 2025). These methods offer fine-grained control but require annotated training data and still struggle with precise timing and occurrence control. The second is spontaneous-style TTS, which predicts NVs directly from contextual cues without requiring aligned annotations (Li et al. 2024a; 2noise 2024). While more flexible and easier to scale, such models often rely on proprietary data or suffer from limited data quality, hindering reproducibility and generalization.

Beyond speech synthesis, understanding non-verbal vocalizations is equally important for building truly human-like speech interaction systems. Traditionally, speech processing systems have focused on single-domain tasks such as automatic speech recognition (ASR) or sound event

*Work conducted when the first author was an intern at ModelBest.

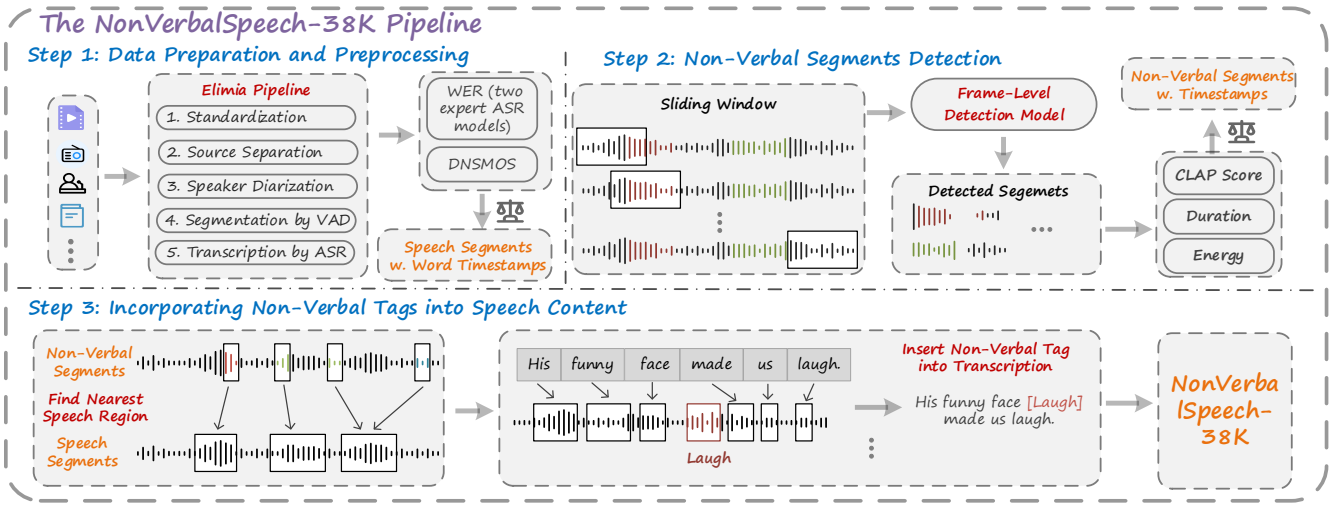

Figure 2: The proposed *NonVerbalSpeech-38K* data processing pipeline. Colors other than black in the waveform indicate non-verbal vocalizations.

classification independently. However, in real-world scenarios, speech and non-verbal sounds naturally co-occur, driving a shift toward unified models that can handle both simultaneously. Recent speech understanding models (Tang et al. 2024; Chu et al. 2023, 2024; Huang et al. 2025; Wu et al. 2025; Ding et al. 2025) employ multi-task training strategies to jointly perform ASR, sound event classification, and related speech tasks. Although this unified approach integrates multiple tasks within a single framework, effectively combining these capabilities—particularly transcribing non-verbal information embedded within speech—remains challenging (Shione, Wakabayashi, and Kitaoka 2023b,a). Progress is further limited by the scarcity of large-scale, fine-grained annotated datasets.

As outlined above, the development of both non-verbal speech synthesis and understanding remains fundamentally hindered by the scarcity of publicly available, large-scale, diverse, and well-annotated datasets. To address this, several efforts have attempted to incorporate non-verbal vocalizations (NVs) into speech datasets (Xin, Takamichi, and Saruwatari 2024; Xin et al. 2024; Wang et al. 2025; Nguyen et al. 2023). However, existing datasets often face significant limitations. Some works rely on manual annotation or human recording, which severely restricts speaker diversity and the scale of the dataset. Other approaches use rule-based data augmentation, which may lead to issues with speech naturalness and timbre inconsistency. Additionally, recent NonverbalTTS (Borisov, Spirin, and Diatlova 2025) has explored semi-supervised pipeline to increase scale, but the resulting dataset is still limited in size and coverage.

To address these limitations, we present **NonVerbalSpeech-38K**, a large-scale dataset of naturally occurring non-verbal vocal expressions with rich text annotations. The data is collected from diverse real-world sources, including animations, movies, variety shows, and radio dramas (Table 3). To ensure scalability, we develop an automatic annotation pipeline for detecting non-verbal events and integrating them with speech transcriptions (Figure 2). The resulting dataset contains 38,718 samples (approximately 131 hours), covering 10 types of non-verbal tags such as cough and laugh (Figure 4). Table 1 highlights the key differences between our dataset and existing resources. We validate the effectiveness of our dataset by fine-tuning two representative models: F5-TTS (Chen et al. 2024) for non-verbal TTS, and Qwen-Audio2 (Chu et al. 2024) for end-to-end captioning that incorporates both verbal and non-verbal cues.

Our main contributions are summarized as follows:

- We design a scalable pipeline for accurate detection and labeling of non-verbal expressions from diverse real-world audio.
- We release a large-scale annotated dataset to facilitate research on non-verbal TTS and non-verbal speech understanding.
- We fine-tune F5-TTS to demonstrate the dataset's utility for non-verbal speech synthesis.
- We fine-tune Qwen2-Audio to investigate joint modeling of speech transcription and non-verbal tag prediction within an end-to-end captioning framework.

2 Related Work

2.1 Speech Dataset with NVs

Recent speech datasets have begun to include non-verbal annotations. For example, JNV corpus (Xin, Takamichi, and Saruwatari 2024) constructs a Japanese NVs dataset from 4 speakers, covering 6 emotions. Then, JVNV (Xin et al. 2024) further uses these NV phrases to create emotional scripts with ChatGPT (Hurst et al. 2024). Besides, Expresso (Nguyen et al. 2023) captures non-verbal vocalizations from natural expressive speech. It incorporates NVs as

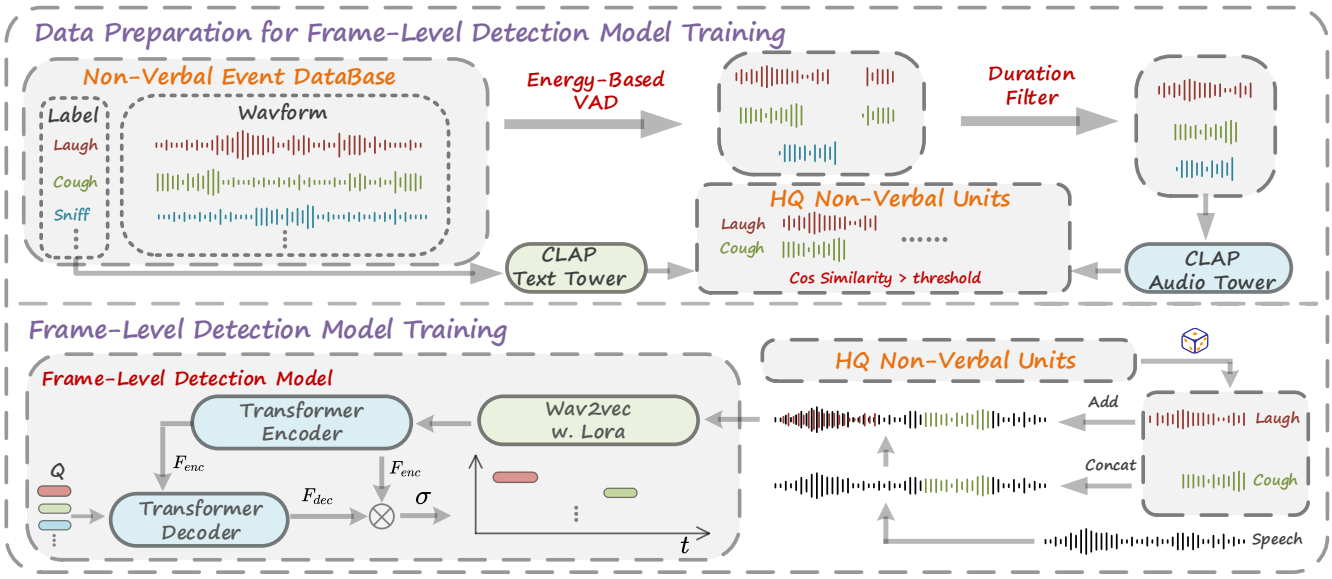

Figure 3: Data preparation and frame-level detection model training.

one of 26 expressive styles, covering a broad range of vocal behaviors that are inherently difficult to transcribe. Despite their valuable resources for non-verbal speech, their limited speakers count severely restricts data diversity. To broaden coverage of NVs and improve diversity, Capspeech-SE (Wang et al. 2025) constructs a non-verbal speech dataset via a rule-based method that integrates multi-speaker speech with sound effect clips. While this creates a large-scale and diverse dataset, it suffers from limitations in timbre consistency and speech naturalness, which hinder its suitability for TTS tasks. Extending previous efforts, NonverbalTTS (Borisov, Spirin, and Diatlova 2025) creates a 17-hour dataset by augmenting VoxCeleb (Nagrani, Chung, and Zisserman 2017; Chung, Nagrani, and Zisserman 2018) and Expresso (Nguyen et al. 2023) with diverse NVs and emotional annotations through a semi-automated pipeline. Although derived from natural speech, it remains limited in scale, and its pipeline lacks scalability. A comparison of the recently proposed datasets Capspeech (Wang et al. 2025) and NonverbalTTS (Borisov, Spirin, and Diatlova 2025) with our dataset is summarized in Table 1.

2.2 Speech Synthesis with NVs

Research on speech synthesis with NVs falls into two main approaches. The first is non-verbal tag-controlled TTS, which inserts explicit tags into the input text to specify NV type and position, such as `[laughing]` or `[breath]`. By training with NV-labeled speech data, FireRedTTS (Guo et al. 2024) and CosyVoice (Du et al. 2024b) have successfully generated natural speech with non-verbal vocalizations. Dia (Nari-labs 2025) further extends this capability to double-speaker dialogue scenarios. The second category is spontaneous-style TTS, which predicts non-verbal vocalizations directly from contextual cues, thus eliminating the need for explicit annotations. Examples include SponTTS (Li et al. 2024a) and ChatTTS (2noise 2024). However, these models often depend on proprietary datasets or are limited by the scale and quality of publicly available data, restricting their generalization and further development.

2.3 Speech Understanding with NVs

Traditional speech processing systems typically focused on single-domain tasks. Nowadays, the current trend has shifted towards building unified speech understanding models that employ multi-task training strategies to jointly address multi-dimension recognition (Tang et al. 2024; Chu et al. 2023, 2024; Huang et al. 2025; Wu et al. 2025). While this unified approach integrates multiple tasks within a single framework, effectively combining these capabilities—especially the transcription of non-verbal information embedded within spoken content—remains challenging (Shione, Wakabayashi, and Kitaoka 2023b,a).

3 Methodology

This section presents our data processing pipeline, *NonVerbalSpeech-38K Pipeline*, as illustrated in Figure 2. It consists of three main stages: Data Preparation and Preprocessing, Non-Verbal Segment Detection, and Integration of Non-Verbal Tags into Speech Content. Among these, Non-Verbal Segment Detection serves as the core component of the pipeline. We detail each step below.

3.1 Data Preparation and Preprocessing

As shown in Figure 2 (Step 1), we collect about 33,252 hours of data from diverse sources such as radio dramas, cartoons, and comedy sketches (Table 3). Using the Emilia-Pipeline (He et al. 2024), we preprocess the data with standardization, source separation, and speaker diarization to extract single-speaker segments. Voice activity detection (VAD)

Dataset	Language	#Samples	Source	Construction Pipeline	# Non-Verbal Tags
NonVerbalTTS	EN	∼6K	VoxCeleb and Expresso	Semi-automatic	10
Capspeech	EN	∼37K	LibriTTS-R	Data augmentation	3
NonVerbalSpeech-38K (Ours)	EN, ZH	∼38K	In-the-wild	Automatic	10

Table 1: Comparison of non-verbal speech datasets in terms of language coverage, size, source, construction pipeline, and supported non-verbal tag types. For Capspeech, we include only the samples containing any of the six target non-verbal tags, as shown in Table 5.

Category	Count	Category	Count
Laughing	11,441	Coughing	9,105
Breath	4,851	Sniff	4,564
Crying	4,369	Throat clearing	4,235
Sigh	2,069	Snore	731
Gasp	333	Yawn	266

Table 2: Statistics of non-verbal segments for training the frame-level detection model.

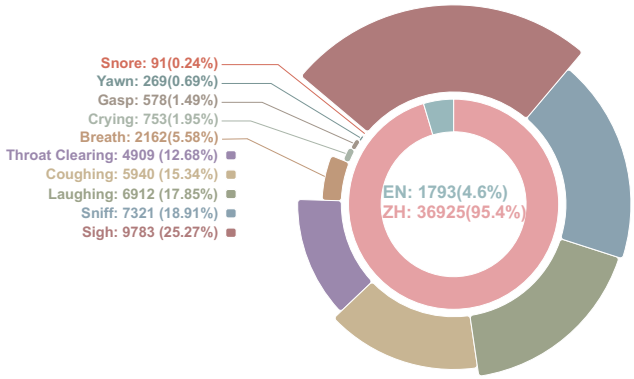

Figure 4: Language and non-verbal label distribution in the proposed NonVerbalSpeech-38K dataset.

removes silence, followed by transcription with Whisper-large-v3 (Radford et al. 2023) and Paraformer-zh (Gao et al. 2022). Segments with WER $>$ 20% or DNSMOS $<$ 1.0 are filtered out to ensure high-quality speech. This process also produces segment and word-level timestamps for further processing.

3.2 Non-Verbal Segments Detection

We train a frame-level detection model to predict precise timestamps of non-verbal events, crucial for embedding non-verbal tags in transcriptions. We first describe the training procedure, then explain its application in the pipeline.

Model Architecture Inspired by T-UAED (Jiang et al. 2025), we propose a frame-level detection model to localize non-verbal events, as illustrated in Figure 3. Specifically, we employ a pretrained wav2vec 2.0 model (Baevski et al. 2020) to extract L-layer features $F_i \in \mathbb{R}^{B\times T\times D}$. To effectively utilize multi-level representations, we apply a learnable weighted sum:

$$F = \sum_{i=1}^{L} \alpha_i F_i, \quad \text{s.t.} \sum \alpha_i = 1,\ \alpha_i \geq 0 \tag{1}$$

The fused representation F is then passed through a Transformer encoder, followed by a Transformer decoder with learnable queries $Q \in \mathbb{R}^{B\times N_q\times D}$:

$$F_{\text{enc}} = \text{Encoder}(F), \quad F_{\text{dec}} = \text{Decoder}(Q, F_{\text{enc}}) \tag{2}$$

To obtain frame-level predictions, we compute the similarity between decoder and encoder outputs:

$$S = \sigma(F_{\text{dec}} F_{\text{enc}}^{\top}), \quad S \in \mathbb{R}^{B\times N_q\times T} \tag{3}$$

Here, $\sigma(\cdot)$ denotes the sigmoid function. Finally, the model is trained using binary cross-entropy loss.

Data Preparation for Model Training To train the frame-level detection model, we construct augmented samples by concatenating or overlaying non-verbal and speech segments, as shown in Figure 3. Non-verbal clips are collected from widely used sound event datasets including VocalSound (Gong, Yu, and Glass 2022), VGGSound (Chen et al. 2020), Nonspeech7k (Rashid, Li, and Du 2023), AudioSet-Strong, and ESC-50 (Piczak 2015), forming our Non-Verbal Event Database. To ensure quality, we apply energy-based VAD to remove silence. Segments with low CLAP scores (below 0.4) or very short durations (less than 0.3 seconds) are discarded. The resulting high-quality clips are referred to as HQ Non-Verbal Units, with statistics summarized in Table 2. Speech data is drawn from the GigaSpeech Medium subset (Chen et al. 2021) and SEED-TTS-Eval (Anastassiou et al. 2024). Only the GigaSpeech test split and the SEED-TTS-Eval set are used for evaluation.

Experimental Setup and Results We set $N_q = 10$, corresponding to the 10 non-verbal types in Table 2. Both encoder and decoder have 6 layers. The model is fine-tuned using LoRA (Hu et al. 2022) for 300 epochs with a batch size of 64 per GPU on 8 H100 GPUs, using a learning rate of 1×10^{-4} with warm-up and cosine annealing. After training, the model achieves a frame-level F1-score of around 91% on three test splits. Additional details are provided in the Appendix.

Non-Verbal Region Detection Using the Trained Frame-Level Detection Model After training the frame-level de-

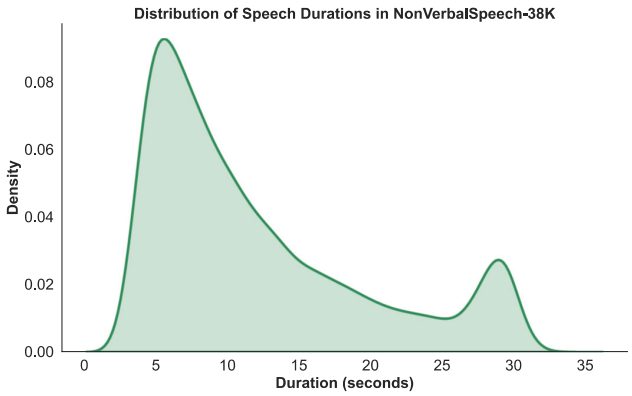

Figure 5: Speech duration distribution of the proposed NonVerbalSpeech-38K dataset.

tection model, we apply it to standardized waveforms using a sliding window strategy, as shown in Figure 2 (Step 2). Standardized waveforms are used because we observe that tools such as Silero-VAD (Team 2024) (used in the Emilia pipeline (He et al. 2024)) often fail to detect non-verbal events occurring at the beginning or end of speech segments. As a result, non-verbal cues at segment boundaries are frequently missed. To provide sufficient context for each frame, we use a window size of 10 seconds and a hop size of 5 seconds.

To improve segment quality, we introduce three filters: first, discard segments shorter than 0.3 seconds (based on empirical observations);); second, compute the maximum energy of each segment and remove those with very low energy, such as subtle breathing sounds that are hard to perceive. In our pipeline, the energy threshold is set to $-35\,\mathrm{dB}$. Third, remove segments with a CLAP score (Elizalde et al. 2023) less than 0.3, as a higher CLAP score indicates that the sound truly corresponds to non-verbal events.

3.3 Incorporating Non-Verbal Tags into Speech Content

After filtering out low-quality candidates, we match each non-verbal segment with the nearest speech region obtained from preprocessing. It should be noted that non-verbal events generally do not lie entirely within speech regions. Therefore, any non-verbal segment located more than 1 second away from the nearest speech region is discarded. The pseudocode for identifying the nearest speech region is provided in the Appendix.

After this, we obtain high-quality speech segments with non-verbal events occurring in the current segment. The available information includes specific non-verbal labels with their start and end timestamps, as well as speech transcriptions with word-level timestamps. To support non-verbal speech generation and understanding, it is necessary to integrate the textual content with non-verbal tags. As shown in Step 3 of Figure 2, for a speech segment with the transcription "`His funny face made us laugh`", there exists a laughter sound right after the utterance "`His funny face`". By aligning the timestamps of the laughter and the speech, we aim to create a final combined transcript like: "`His funny face [laugh] made us laugh.`". The related algorithm is provided in the Appendix.

Source	Radio Dramas	Comedy Sketches	Cartoon	Variety Shows	Short Plays
Crawled (hrs)	17,400	7,200	3,600	1,400	1,200
NVS-38K (#)	21,668	7,273	5,019	1,217	809
Source	Speeches	Documentaries	Movies	Audiobooks	Toy Unboxing
Crawled (hrs)	1,079	600	500	263	9
NVS-38K (#)	158	105	1,090	1,375	4

Table 3: "Crawled (hrs)" shows the duration of the original crawled data; "NVS-38K (#)" shows the final sample counts in the proposed NonVerbalSpeech-38K dataset.

4 The NonVerbalSpeech-38K Dataset

Using the proposed NonVerbalSpeech-38K pipeline, we construct a dataset from diverse online speech sources, covering a wide range of natural speaking styles and rich non-verbal vocalizations. The raw data statistics are shown in Table 3. Radio dramas, comedy sketches, and cartoons contribute the most data, while movies and audiobooks yield the highest processing conversion rates, suggesting richness in non-verbal content. The initial release contains $\sim$131 hours of English and Chinese speech, totaling about 38,000 samples. As shown in Figure 4, Chinese samples dominate due to the composition of the source data. Interestingly, the frame-level detection model trained only on English still generalizes well to Chinese, indicating strong cross-lingual scalability. Figure 4 also shows the distribution of non-verbal tags. Frequent types such as sigh, sniff, laughing, and coughing dominate, reflecting their prevalence in daily speech. Less frequent types (e.g., crying, yawn) are also included. The resulting durations mostly fall within the range of 3 to 30 seconds (Figure 5). More information is available in the Appendix.

5 Experiments

We evaluate the effectiveness of our dataset through two downstream tasks: non-verbal speech generation and end-to-end non-verbal speech understanding. The details are as follows.

5.1 Experimental Setup for Non-Verbal Speech Generation

Implementation Details In the non-verbal speech generation task, we use F5-TTS (Chen et al. 2024) as the base model. To support non-verbal control, we expand the vocabulary of F5-TTS by adding non-verbal tags such as `[laughing]`, `[coughing]`. We fine-tune F5-TTS on three datasets: CapSpeech (Wang et al. 2025), NonVerbalTTS (Borisov, Spirin, and Diatlova 2025), and our pro-

Model	CLAP Score ↑		WER/CER (%) ↓		DNSMOS ↑		SSIM ↑		QMOS ↑		IMOS ↑	
	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH
Baseline Models												
CosyVoice2	0.090	0.074	3.930	5.023	3.240	3.288	0.514	0.671	4.09	4.06	2.29	2.54
Dia	0.182	0.097	15.687	171.882	2.749	2.528	0.431	0.319	3.33	–	3.51	–
F5-TTS and Variants												
F5-TTS	0.044	0.036	5.325	6.188	3.046	3.200	0.673	0.721	3.61	3.22	–	–
+ CapSpeech	**0.147**	**0.119**	10.539	16.137	2.889	3.048	<u>0.563</u>	<u>0.661</u>	2.75	2.01	2.02	2.11
+ NVTTS	0.103	0.076	**8.013**	<u>15.808</u>	**3.069**	**3.149**	**0.587**	0.648	**3.63**	<u>2.53</u>	**3.25**	<u>2.80</u>
+ NVS (Ours)	<u>0.104</u>	<u>0.102</u>	<u>10.132</u>	**7.341**	<u>3.033</u>	<u>3.115</u>	0.544	**0.671**	<u>3.43</u>	**3.63**	<u>3.22</u>	**3.44**
	0.137	*0.146*	–	–	–	–	–	–	–	–	–	–

Table 4: Evaluation results of non-verbal speech generation on English (EN) and Chinese (ZH) datasets. The best and second-best values in each column are highlighted in **bold** and <u>underline</u>, respectively. CLAP Score and IMOS are computed only on tags supported by each model or dataset, as specified in Table 5. The last row presents the CLAP Score of F5-TTS fine-tuned on our proposed dataset, evaluated solely on tags supported by CapSpeech. Here, NVTTS stands for NonVerbalTTS, and NVS stands for NonVerbalSpeech-38K.

Label	Datasets			Models	
	Cap	NVTTS	NVS38K	CV2	Dia
Coughing	**Y**	**Y**	**Y**	**Y**	**Y**
Laughing	**Y**	**Y**	**Y**	**Y**	**Y**
Sigh	*N*	**Y**	**Y**	*N*	**Y**
Sniff	*N*	**Y**	**Y**	*N*	**Y**
Throat Clearing	*N*	**Y**	**Y**	*N*	**Y**
Breath	**Y**	**Y**	**Y**	**Y**	*N*

Table 5: Non-verbal label support across datasets and models. Cap: CapSpeech, NVTTS: NonVerbalTTS, NVS38K: NonVerbalSpeech-38K, CV2: CosyVoice2.

posed NonVerbalSpeech-38K dataset. More detailed information can be found in the Appendix.

Baselines In addition to fine-tuning F5-TTS on different datasets, we compare it with two strong baselines: CosyVoice2 (Du et al. 2024b) and Dia (Nari-labs 2025), both of which support non-verbal control and generation (Table 5).

Evaluation Due to differences in the non-verbal tags supported by each model and dataset (Table 5), we select six common non-verbal phenomena for evaluation: `[laughing]`, `[sniff]`, `[breath]`, `[coughing]`, `[throatclearing]`, and `[sigh]`. To evaluate the non-verbal speech generation capabilities of different models, we reuse the prompt speech provided by the seed-tts-eval framework (Anastassiou et al. 2024). For the text inputs, we generate tagged sentences using Genimi (Team et al. 2023), GPT-4o (Hurst et al. 2024). For each tag, we construct 175 sentences in both Chinese and English, resulting in a total of 1,050 sentences per language.

To evaluate system performance, we use standard metrics including DNSMOS, SSIM, and WER/CER to assess speech quality, speaker similarity, and intelligibility. Additionally, we introduce the CLAP Score to measure the presence of intended non-verbal phenomena in synthesized speech. For subjective evaluation, 20 participants rated the samples on QMOS (for overall speech quality) and IMOS (for accuracy in rendering both verbal content and non-verbal tags). See Appendix for scoring details.

5.2 Experimental Setup for End-to-End Non-Verbal Speech Understaning

Implementation Details In this task, we use Qwen2-Audio (Chu et al. 2024) as our base model. Qwen2-Audio is an advanced audio-language model capable of processing various audio inputs and generating either analytical results or textual responses. However, it can only classify non-verbal events or transcribe speech, and does not support integrating non-verbal tags into the transcription. To enable this capability, we introduce special tokens corresponding to non-verbal tags. We then fine-tune Qwen2-Audio on the CapSpeech (Wang et al. 2025) and NonVerbalSpeech-38K datasets to equip the model with the ability to understand and generate these tags. The details of the fine-tuning process can be found in the Appendix.

Evaluation To evaluate the model's ability to integrate non-verbal tags into transcriptions, we conduct experiments on NonVerbalTTS (Borisov, Spirin, and Diatlova 2025), a high-quality, human-annotated real-world dataset specifically designed for non-verbal speech modeling. We select a subset of 754 samples containing only the three tags: `[laughing]`, `[breath]`, and `[coughing]`. For simplicity in evaluation, each utterance contains one tag. The number of samples per tag is listed in the Appendix.

In addition to the standard word error rate (WER), we use three tag-level metrics evaluating type and position accuracy: (1) **F1-Score** for tag correctness; and (2) **Tag Posi-**

Model	WER Details			F1-Score ↑	TPD ↓	NTD ↓
	WER (%)	WER $<$ 50% (%)	# WER $<$ 50%			
Whisper-Large-V3	17.542	8.693	731 / 754	–	–	–
Qwen2-Audio	33.087	11.603	620 / 754	–	–	–
+ CapSpeech	23.806	16.508	709 / 754	0.598	**4.232**	**0.172**
+ NonVerbalSpeech-38K (Ours)	**20.215**	**14.445**	727 / 754	**0.661**	6.223	0.284

Table 6: Evaluation results for end-to-end non-verbal speech understanding. WER $<$ 50% refers to WER calculated after removing samples with extremely high error rates (WER $>$ 50%). # WER $<$ 50% denotes the count of samples with WER below 50% over the total number of test samples.

tion Distance (TPD) and its normalized version **Normalized Tag Distance (NTD)**. TPD and NTD are computed after aligning predicted and reference tags using Levenshtein edit distance, where

$$\text{TPD} = \frac{1}{n}\sum_{i=1}^{n}|p_i - g_i|, \quad \text{NTD} = \frac{\text{TPD}}{L},$$

with p_i, g_i the positions of matched tags and L the reference length. Lower TPD and NTD indicate better positional alignment. Detailed definitions are provided in the Appendix.

6 Experimental Results

6.1 Experimental Results on Non-Verbal TTS

Table 4 summarizes non-verbal speech generation results for English (EN) and Chinese (ZH). Among baselines, CosyVoice2 excels in standard speech metrics (lowest WER/CER, highest DNSMOS and QMOS) but shows limited non-verbal controllability (lower CLAP-Score and IMOS). Dia performs best on English non-verbal metrics (CLAP and IMOS) but suffers from poor stability and high failure rates, especially unsupported in Chinese, leading to exclusion of its Chinese subjective scores. F5-TTS serves as a strong baseline with competitive performance metrics, achieving the highest SSIM. However, the pretrained F5-TTS does not support non-verbal TTS, and its CLAP-Score is reported for reference only. Our fine-tuned F5-TTS variant shows superior Chinese performance, achieving top subjective scores (QMOS, IMOS) and strong objective metrics, demonstrating competitive intelligibility and naturalness, as well as improved non-verbal controllability. CapSpeech and NVTTS underperform in Chinese due to English-only training data, reflected in lower QMOS, IMOS, SSIM and higher CER. On English tests, our NVS matches NVTTS on IMOS and CLAP-Score despite limited English data (4.6%). CapSpeech attains high CLAP but low DNSMOS and QMOS, likely due to synthetic non-verbal integration. When restricting the CLAP-Score evaluation to tags supported by CapSpeech (Table 5), as shown in the last line of Table 4, the model trained on the proposed dataset outperforms CapSpeech.

Overall, the diversity and naturalness of our dataset contribute to the competitive intelligibility, naturalness, and controllability achieved in multilingual non-verbal speech synthesis.

6.2 Experimental Results on End-to-End Non-Verbal Speech Understaning

Table 6 compares end-to-end non-verbal speech understanding performance between Whisper-Large-V3, Qwen2-Audio, and its fine-tuned variants. Whisper-Large-V3 achieves the best transcription accuracy (lowest WER). The original Qwen2-Audio shows much higher WER and does not support non-verbal tag integration, so tag-related metrics are unavailable. Fine-tuning Qwen2-Audio on our dataset significantly improves both transcription and non-verbal tag recognition. Our fine-tuned model attains higher F1-Score than the CapSpeech variant, indicating better tag detection accuracy. Although its TPD and NTD are slightly worse due to ASR transcription misalignment during tag insertion (Figure 2, Step 3), the results show that training on diverse real data improves robustness compared to CapSpeech, which relies on fixed sound event sets and limited speech corpora.

7 Discussion

Finally, we discuss the scalability of our pipeline, which is reflected in two main aspects:

- The frame-level detection model is extensible to general sounds beyond non-verbal expressions by increasing the number of query tokens N_q. With CLAP (Elizalde et al. 2023) text embeddings as input, it also supports description-based sound and effect detection for finer-grained recognition.
- Unlike prior works such as NonVerbalTTS (Borisov, Spirin, and Diatlova 2025), our pipeline is fully automated, requiring no manual verification. Compared to Capspeech (Wang et al. 2025), it yields more natural data distributions not constrained by reading-style corpora like LibriTTS (Zen et al. 2019).

8 Conclusion

In this work, we present **NonVerbalSpeech-38K**, a large, diverse dataset for advancing non-verbal TTS and end-to-end speech captioning. By collecting naturally expressive audio from diverse media and developing an automated annotation pipeline, it enhances current speech synthesis and understanding systems in non-verbal generation and understanding, respectively—with its effectiveness validated by experiments using state-of-the-art models.

References

2noise. 2024. chatts. https://github.com/2noise/ChatTTS.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anastassiou, P.; Chen, J.; Chen, J.; Chen, Y.; Chen, Z.; Chen, Z.; Cong, J.; Deng, L.; Ding, C.; Gao, L.; et al. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Borisov, M.; Spirin, E.; and Diatlova, D. 2025. NonverbalTTS: A Public English Corpus of Text-Aligned Nonverbal Vocalizations with Emotion Annotations for Text-to-Speech. *arXiv preprint arXiv:2507.13155*.

Chen, G.; Chai, S.; Wang, G.-B.; Du, J.; Zhang, W.-Q.; Weng, C.; Su, D.; Povey, D.; Trmal, J.; Zhang, J.; et al. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Proc. Interspeech 2021*, 3670–3674.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725. IEEE.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518.

Chen, Y.; Niu, Z.; Ma, Z.; Deng, K.; Wang, C.; Zhao, J.; Yu, K.; and Chen, X. 2024. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. *arXiv preprint arXiv:2311.07919*.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. VoxCeleb2: Deep Speaker Recognition. In *Interspeech 2018*, 1086–1090.

Clancey, W. J. 1979. *Transfer of Rule-Based Expertise through a Tutorial Dialogue*. Ph.D. diss., Dept. of Computer Science, Stanford Univ., Stanford, Calif.

Clancey, W. J. 1983. Communication, Simulation, and Intelligent Agents: Implications of Personal Intelligent Machines for Medical Education. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83)*, 556–560. Menlo Park, Calif: IJCAI Organization.

Clancey, W. J. 1984. Classification Problem Solving. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, 45–54. Menlo Park, Calif.: AAAI Press.

Clancey, W. J. 2021. The Engineering of Qualitative Models. Forthcoming.

Cortes, D. S.; Tornberg, C.; Bänziger, T.; Elfenbein, H. A.; Fischer, H.; and Laukka, P. 2021. Effects of aging on emotion recognition from dynamic multimodal expressions and vocalizations. *Scientific reports*, 11(1): 2647.

Cowen, A. S.; Elfenbein, H. A.; Laukka, P.; and Keltner, D. 2019. Mapping 24 emotions conveyed by brief human vocalization. *American psychologist*, 74(6): 698.

Défossez, A.; Mazaré, L.; Orsini, M.; Royer, A.; Pérez, P.; Jégou, H.; Grave, E.; and Zeghidour, N. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Ding, D.; Ju, Z.; Leng, Y.; Liu, S.; Liu, T.; Shang, Z.; Shen, K.; Song, W.; Tan, X.; Tang, H.; et al. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Du, Z.; Chen, Q.; Zhang, S.; Hu, K.; Lu, H.; Yang, Y.; Hu, H.; Zheng, S.; Gu, Y.; Ma, Z.; et al. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Du, Z.; Gao, C.; Wang, Y.; Yu, F.; Zhao, T.; Wang, H.; Lv, X.; Wang, H.; Ni, C.; Shi, X.; et al. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Du, Z.; Wang, Y.; Chen, Q.; Shi, X.; Lv, X.; Zhao, T.; Gao, Z.; Yang, Y.; Gao, C.; Wang, H.; et al. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Elizalde, B.; Deshmukh, S.; Al Ismail, M.; and Wang, H. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Engelmore, R.; and Morgan, A., eds. 1986. *Blackboard Systems*. Reading, Mass.: Addison-Wesley.

Fang, Q.; Guo, S.; Zhou, Y.; Ma, Z.; Zhang, S.; and Feng, Y. ???? LLaMA-Omni: Seamless Speech Interaction with Large Language Models. In *The Thirteenth International Conference on Learning Representations*.

Fang, Q.; Guo, S.; Zhou, Y.; Ma, Z.; Zhang, S.; and Feng, Y. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Gao, Z.; Zhang, S.; McLoughlin, I.; and Yan, Z. 2022. Paraformer: Fast and Accurate Parallel Transformer for Non-autoregressive End-to-End Speech Recognition. In *Proc. Interspeech 2022*, 2063–2067.

Gong, Y.; Yu, J.; and Glass, J. 2022. Vocalsound: A Dataset for Improving Human Vocal Sounds Recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 151–155.

Guo, H.-H.; Hu, Y.; Liu, K.; Shen, F.-Y.; Tang, X.; Wu, Y.-C.; Xie, F.-L.; Xie, K.; and Xu, K.-T. 2024. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. *arXiv preprint arXiv:2409.03283*.

Hasling, D. W.; Clancey, W. J.; and Rennels, G. 1984. Strategic explanations for a diagnostic consultation system. *International Journal of Man-Machine Studies*, 20(1): 3–19.

Hasling, D. W.; Clancey, W. J.; Rennels, G. R.; and Test, T. 1983. Strategic Explanations in Consultation—Duplicate. *The International Journal of Man-Machine Studies*, 20(1): 3–19.

He, H.; Shang, Z.; Wang, C.; Li, X.; Gu, Y.; Hua, H.; Liu, L.; Yang, C.; Li, J.; Shi, P.; et al. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, 885–890. IEEE.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, A.; Wu, B.; Wang, B.; Yan, C.; Hu, C.; Feng, C.; Tian, F.; Shen, F.; Li, J.; Chen, M.; Liu, P.; Miao, R.; You, W.; Chen, X.; Yang, X.; Huang, Y.; Zhang, Y.; Gong, Z.; Zhang, Z.; Li, B.; Wan, C.; Hu, H.; Ming, R.; Yuan, S.; Zhang, X.; Zhou, Y.; Li, B.; Ma, B.; An, K.; Ji, W.; Li, W.; Wen, X.; Ma, Y.; Liang, Y.; Mou, Y.; Ahmidi, B.; Wang, B.; Li, B.; Miao, C.; Xu, C.; Feng, C.; Wang, C.; Shi, D.; Sun, D.; Hu, D.; Sai, D.; Liu, E.; Huang, G.; Yan, G.; Wang, H.; Jia, H.; Zhang, H.; Gong, J.; Wu, J.; Liu, J.; Sun, J.; Zhen, J.; Feng, J.; Wu, J.; Wu, J.; Yang, J.; Wang, J.; Zhang, J.; Lin, J.; Li, K.; Xia, L.; Zhou, L.; Gu, L.; Chen, M.; Wu, M.; Li, M.; Li, M.; Liang, M.; Wang, N.; Hao, N.; Wu, Q.; Tan, Q.; Pang, S.; Yang, S.; Gao, S.; Liu, S.; Liu, S.; Cao, T.; Wang, T.; Deng, W.; He, W.; Sun, W.; Han, X.; Deng, X.; Liu, X.; Zhao, X.; Wei, Y.; Yu, Y.; Cao, Y.; Li, Y.; Ma, Y.; Xu, Y.; Shi, Y.; Wang, Y.; Zhong, Y.; Luo, Y.; Lu, Y.; Yin, Y.; Yan, Y.; Yang, Y.; Xie, Z.; Ge, Z.; Sun, Z.; Huang, Z.; Chang, Z.; Yang, Z.; Zhang, Z.; Jiao, B.; Jiang, D.; Shum, H.-Y.; Chen, J.; Li, J.; Zhou, S.; Zhang, X.; Zhang, X.; and Zhu, Y. 2025. Step-Audio: Unified Understanding and Generation in Intelligent Speech Interaction. arXiv:2502.11946.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jiang, Y.; Tao, R.; Huang, W.; Chen, Q.; and Wang, W. 2025. Unified audio event detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Ju, Z.; Yang, D.; Yu, J.; Shen, K.; Leng, Y.; Wang, Z.; Tan, X.; Zhou, X.; Qin, T.; and Li, X. 2025. MoonCast: High-quality zero-shot podcast generation. *arXiv preprint arXiv:2503.14345*.

Kamiloğlu, R. G.; Fischer, A. H.; and Sauter, D. A. 2020. Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic bulletin & review*, 27(2): 237–265.

Koizumi, Y.; Zen, H.; Karita, S.; Ding, Y.; Yatabe, K.; Morioka, N.; Bacchiani, M.; Zhang, Y.; Han, W.; and Bapna, A. 2023. LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus. In *Interspeech 2023*, 5496–5500.

Lee, S.-g.; Ping, W.; Ginsburg, B.; Catanzaro, B.; and Yoon, S. 2022. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*.

Li, H.; Zhu, X.; Xue, L.; Song, Y.; Chen, Y.; and Xie, L. 2024a. Spontts: modeling and transferring spontaneous style for tts. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12171–12175. IEEE.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, W.; Lei, S.; Huang, Q.; Zhou, Y.; Wu, Z.; Kang, S.; and Meng, H. 2023b. Towards Spontaneous Style Modeling with Semi-supervised Pre-training for Conversational Text-to-Speech Synthesis. In *Proc. Interspeech 2023*, 3377–3381.

Li, W.; Yang, P.; Zhong, Y.; Zhou, Y.; Wang, Z.; Wu, Z.; Wu, X.; and Meng, H. 2024b. Spontaneous Style Text-to-Speech Synthesis with Controllable Spontaneous Behaviors Based on Language Models. In *Proc. Interspeech 2024*, 1785–1789.

Lima, C. F.; Alves, T.; Scott, S. K.; and Castro, S. L. 2014. In the ear of the beholder: how age shapes emotion processing in nonverbal vocalizations. *Emotion*, 14(1): 145.

McAuliffe, M.; Socolof, M.; Mihuc, S.; Wagner, M.; and Sonderegger, M. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, 498–502.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Interspeech 2017*, 2616–2620.

Nari-labs. 2025. Dia. https://github.com/nari-labs/dia.

NASA. 2015. Pluto: The 'Other' Red Planet. https://www.nasa.gov/nh/pluto-the-other-red-planet. Accessed: 2018-12-06.

Nguyen, T. A.; Hsu, W.-N.; d'Avirro, A.; Shi, B.; Gat, I.; Fazel-Zarani, M.; Remez, T.; Copet, J.; Synnaeve, G.; Hassid, M.; et al. 2023. Expresso: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis. In *INTERSPEECH 2023-24th Annual Conference of the International Speech Communication Association*, 4823–4827. ISCA.

Piczak, K. J. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, 1015–1018.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Rashid, M. M.; Li, G.; and Du, C. 2023. Nonspeech7k dataset: Classification and analysis of human non-speech sound. *IET Signal Processing*, 17(6): e12233.

Rice, J. 1986. Poligon: A System for Parallel Problem Solving. Technical Report KSL-86-19, Dept. of Computer Science, Stanford Univ.

Robinson, A. L. 1980a. New Ways to Make Microcircuits Smaller. *Science*, 208(4447): 1019–1022.

Robinson, A. L. 1980b. New Ways to Make Microcircuits Smaller—Duplicate Entry. *Science*, 208: 1019–1026.

Schmid, F.; Morocutti, T.; Foscarin, F.; Schlüter, J.; Primus, P.; and Widmer, G. 2025. Effective pre-training of audio transformers for sound event detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Shione, N.; Wakabayashi, Y.; and Kitaoka, N. 2023a. Automatic Speech Recognition Using Linguistic and Verbal/Non-Verbal Information. In *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, 1–6.

Shione, N.; Wakabayashi, Y.; and Kitaoka, N. 2023b. Construction of Automatic Speech Recognition Model that Recognizes Linguistic Information and Verbal/Non-verbal Phenomena. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2306–2311.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; MA, Z.; and Zhang, C. 2024. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, S. 2024. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. https://github.com/snakers4/silero-vad.

Truong, K. P.; and Van Leeuwen, D. A. 2007. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2): 144–158.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017a. Attention Is All You Need. arXiv:1706.03762.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017b. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Wang, H.; Hai, J.; Chong, D.; Thakkar, K.; Feng, T.; Yang, D.; Lee, J.; Velazquez, L. M.; Villalba, J.; Qin, Z.; et al. 2025. CapSpeech: Enabling Downstream Applications in Style-Captioned Text-to-Speech. *arXiv preprint arXiv:2506.02863*.

Wang, X.; Li, Y.; Fu, C.; Shen, Y.; Xie, L.; Li, K.; Sun, X.; and Ma, L. 2024. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *arXiv preprint arXiv:2411.00774*.

Wu, B.; Yan, C.; Hu, C.; Yi, C.; Feng, C.; Tian, F.; Shen, F.; Yu, G.; Zhang, H.; Li, J.; Chen, M.; Liu, P.; You, W.; Zhang, X. T.; Li, X.; Yang, X.; Deng, Y.; Huang, Y.; Li, Y.; Zhang, Y.; You, Z.; Li, B.; Wan, C.; Hu, H.; Zhen, J.; Chen, S.; Yuan, S.; Zhang, X.; Jiang, Y.; Zhou, Y.; Yang, Y.; Li, B.; Ma, B.; Song, C.; Pang, D.; Hu, G.; Sun, H.; An, K.; Wang, N.; Gao, S.; Ji, W.; Li, W.; Sun, W.; Wen, X.; Ren, Y.; Ma, Y.; Lu, Y.; Wang, B.; Li, B.; Miao, C.; Liu, C.; Xu, C.; Shi, D.; Hu, D.; Wu, D.; Liu, E.; Huang, G.; Yan, G.; Zhang, H.; Nie, H.; Jia, H.; Zhou, H.; Sun, J.; Wu, J.; Wu, J.; Yang, J.; Yang, J.; Lin, J.; Li, K.; Yang, L.; Shi, L.; Zhou, L.; Gu, L.; Li, M.; Li, M.; Li, M.; Wu, N.; Han, Q.; Tan, Q.; Pang, S.; Fan, S.; Liu, S.; Cao, T.; Lu, W.; He, W.; Xie, W.; Zhao, X.; Li, X.; Yu, Y.; Yang, Y.; Liu, Y.; Lu, Y.; Wang, Y.; Ding, Y.; Liang, Y.; Lu, Y.; Luo, Y.; Yin, Y.; Zhan, Y.; Zhang, Y.; Yang, Z.; Zhang, Z.; Jiao, B.; Jiang, D.; Shum, H.-Y.; Chen, J.; Li, J.; Zhang, X.; and Zhu, Y. 2025. Step-Audio 2 Technical Report. arXiv:2507.16632.

Xie, Z.; and Wu, C. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.

Xin, D.; Jiang, J.; Takamichi, S.; Saito, Y.; Aizawa, A.; and Saruwatari, H. 2024. JVNV: A corpus of japanese emotional speech with verbal content and nonverbal expressions. *IEEE Access*, 12: 19752–19764.

Xin, D.; Takamichi, S.; and Saruwatari, H. 2024. JNV corpus: A corpus of Japanese nonverbal vocalizations with diverse phrases and emotions. *Speech Communication*, 156: 103004.

Yang, Z.; Chen, Y.; Luo, L.; Yang, R.; Ye, L.; Cheng, G.; Xu, J.; Jin, Y.; Zhang, Q.; Zhang, P.; et al. 2022. Open Source MagicData-RAMC: A Rich Annotated Mandarin Conversational (RAMC) Speech Dataset. In *Proc. Interspeech 2022*, 1736–1740.

Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R. J.; Jia, Y.; Chen, Z.; and Wu, Y. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; and Qiu, X. 2023. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 15757–15773.

A Experimental Setup

In this section, we detail the experimental setup for the frame-level detection model, non-verbal speech generation, and non-verbal speech understanding.

Dataset	Laughing	Coughing	Breath	Total
CapSpeech (Train)	19,571	15,174	2,110	36,855
NonVerbalSpeech-38K Subset (Train)	6,912	5,940	2,162	15,014
NonVerbalTTS Subset (Eval)	**367**	**137**	**250**	**754**

Table 7: Counts of non-verbal tags in training and evaluation sets. CapSpeech and a subset of NonVerbalSpeech-38K are used for training, and the NonVerbalTTS subset is used for evaluation.

A.1 Experimental Setup and Results for Frame-level Detection Model

We set N_q to 10, corresponding to the 10 non-verbal phenomena described in Table 2, and the number of layers L for the pre-trained wav2vec 2.0 model (Baevski et al. 2020) is 25. Both the encoder and decoder are configured with 6 layers, where the decoder excludes the self-attention module. The Q embeddings are randomly initialized. We adopt LoRA (Hu et al. 2022) to fine-tune all linear layers within the self-attention modules of the wav2vec 2.0 backbone. The model is trained for 300 epochs with a batch size of 64 per GPU, using 8 NVIDIA H100 GPUs. The learning rate is set to 1×10^{-4} and scheduled with a warm-up phase followed by cosine annealing decay.

After training, the frame-level F1-Score achieves approximately 91% across all three test splits, namely the GigaSpeech test split (Chen et al. 2021), SEED-TTS-Eval Chinese speech (Anastassiou et al. 2024), and SEED-TTS-Eval English speech (Anastassiou et al. 2024). Note that these test sets are also constructed by randomly concatenating or overlaying non-verbal clips with speech segments.

	Aligned by Edit Distance
REF	`[coughing] i just did what **** everybody else does you know after ***** school they kind of look for the next thing and the next thing was you know a levels and`
HYP	`[coughing] i just did what what everybody else does you know after after school he kind of looked for the next thing and the next thing was you know a levels and`

Table 8: Example alignment between reference (REF) and model hypothesis (HYP) with non-verbal tag `[coughing]`. Insertions (e.g., repeated `what`, `after`) and substitutions (e.g., `he` vs. `they`) are observed.

A.2 Experimental Setup for Non-Verbal Speech Generation

Implementation Details In the non-verbal speech generation task, we use F5-TTS (Chen et al. 2024) as the base model. It is a fully non-autoregressive speech synthesis model that adopts a flow-matching method combined with a transformer architecture. We use the pretrained F5-TTS model, which was trained on the Chinese and English subsets of the Emilia dataset (He et al. 2024), and further fine-tune it on our target data. We use BigVGAN-v2 (Lee et al. 2022) as the vocoder. BigVGAN-v2 is trained on datasets that include diverse types of audio, such as speech in multiple languages, environmental sounds, and musical instruments. This diversity is important for non-verbal speech generation, as it helps the vocoder better handle a wide range of acoustic patterns.

To support non-verbal control, we expand the vocabulary of F5-TTS by adding non-verbal tags such as `[laughing]`, `[coughing]`. Inspired by CapSpeech (Wang et al. 2025), we initialize the embeddings of the new tokens using representations from the CLAP (Elizalde et al. 2023) text encoder, which is trained on large-scale audio-text alignment data, including non-verbal audio-text pairs.

We fine-tune F5-TTS on three datasets: CapSpeech (Wang et al. 2025), NonVerbalTTS (Borisov, Spirin, and Diatlova 2025), and our proposed NonVerbalSpeech-38K dataset. To retain the zero-shot synthesis ability of the pretrained model while allowing it to learn non-verbal speech generation, we carefully adjust the total number of training steps according to each dataset. Based on empirical observations, we train both NonVerbalTTS and our proposed NonVerbalSpeech-38K dataset for 400 epochs, while CapSpeech is only trained for 50 epochs. We find that training CapSpeech for 400 epochs significantly degrades intelligibility, making the synthesized speech hard to understand. All models are optimized using the AdamW optimizer with a base learning rate of $1\mathrm{e}{-5}$, a 10% warm-up phase, and a cosine annealing learning rate schedule. For the newly added non-verbal tokens,

Metrics	QMOS: Quality Evaluation	IMOS: Instruction Following Evaluation
Guidance	Please rate each audio sample based on *naturalness* (whether it sounds like real human speech), *sound quality* (whether it is clean and undistorted), and *intelligibility* (how easily the content can be understood). Note: Segments containing only silence should be excluded from the evaluation of *naturalness* and *intelligibility*.	Please rate each audio sample based on how well it follows the given text instructions. For example, given the text: `[sigh] Today's work is not even halfway done yet.`. The TTS system is expected not only to synthesize the text correctly but also to follow the `[sigh]` tag by producing a sigh sound at the corresponding position. All tags used in the evaluation and their meanings are as follows: `[laughing]`: laughing sound, `[coughing]`: coughing sound, `[sigh]`: sighing sound, `[sniff]`: sniffing sound, `[throatclearing]`: throat clearing sound, `[breath]`: breathing sound.
Score	**Criteria**	**Criteria**
5	Excellent naturalness, pure sound quality, and perfectly intelligible content.	Fully follows the text and accurately renders the non-verbal tags, the speech sounds entirely human-like.
4	Good naturalness, mostly clean sound, and clear content.	Fully follows the text, and non-verbal tags are rendered naturally.
3	Moderate naturalness, minor sound artifacts (e.g., slight noise), and mostly intelligible content.	Mostly follows the text; non-verbal tag rendering is generally accurate but lacks naturalness.
2	Poor naturalness, noticeable noise or distortion, and hard-to-understand content.	Slight omission of text, missing or incorrectly rendered tags (e.g., misplaced tag or wrong sound).
1	Very unnatural (e.g., mechanical), heavily distorted, and nearly unintelligible.	Significant text loss or synthesis failure (e.g., unintelligible or missing content).

Table 9: Subjective evaluation criteria for QMOS and IMOS.

a higher learning rate of $1e{-}4$ is applied to encourage faster adaptation. Fine-tuning is performed on 8 NVIDIA H100 GPUs, with a batch size of 38,400 frames per GPU.

Evaluation To evaluate system performance, in addition to standard metrics such as DNSMOS, SSIM (speaker similarity), and WER/CER (Word Error Rate or Character Error Rate) for assessing speech quality, speaker similarity, and intelligibility, respectively, we also introduce the CLAP Score to evaluate whether the synthesized speech contains the intended non-verbal phenomena. Specifically, we use the sentence "`This is a sound of` **`non-verbal-tag.`**" as the text input to the CLAP (Elizalde et al. 2023) text tower, and feed the full synthesized waveform into the CLAP audio tower. For subjective evaluation, we adopt two metrics. The first is QMOS, which assesses the overall quality of the synthesized speech, focusing on intelligibility, audio quality, and naturalness. The second is IMOS, which evaluates whether the audio follows the input text, including non-verbal tags. For example, given the input "`[sigh] The work for today is not even halfway done yet.`", the TTS system is expected not only to correctly synthesize the verbal content but also to emit a sigh sound at the appropriate position as indicated by the `[sigh]` tag. We collected a total of 20 subjective questionnaires. Detailed scoring criteria are provided in the table 9.

A.3 Experimental Setup for End-to-End Non-Verbal Speech Understaning

Implementation Details In this task, we use Qwen2-Audio (Chu et al. 2024) as our base model. The fine-tuning is performed using LoRA, applied to the linear layers within the self-attention modules of the language model. During training, only the LoRA parameters and the newly introduced tokens are updated, while all other model parameters are frozen. For a fair comparison, we use only the subset of NonVerbalSpeech-38K that contains the three non-verbal tags supported by CapSpeech (i.e., `[laughing]`, `[breath]`, and `[coughing]`), the details are shown in Table 7. Both datasets are trained for 5 epochs. All models are trained using a learning rate of 7.5e-5, batch size 12 per GPU, with 10% warmup and cosine learning rate decay, using 8 NVIDIA H100 GPUs.

Evaluation The subset of NonVerbalTTS for evaluation is shown in Table 7. To simplify the evaluation, we include only samples that contain a single non-verbal tag, resulting in 367 `[laughing]`, 137 `[coughing]`, and 1,625 `[breath]` samples. To ensure a balanced tag distribution, 250 samples are randomly selected from the 1,625 available `[breath]` samples. The resulting evaluation set accounts for approximately 5% of the NonVerbalSpeech-38K subset used for training.

Algorithm 1: Find Nearest Speech Region for a Non-Verbal Segment

Require: Speech regions $S = \{s_i\}$; non-verbal segment n
Ensure: Nearest $s \in S$ w.r.t. n or discard n

$(d_{\min}, s^*) \leftarrow (+\infty, \varnothing)$
for $s \in S$ **do**
  $d \leftarrow \text{distance}(n, s)$
  if $n \cap s \neq \varnothing$ **then**
    $s^* \leftarrow s$
    break
  else if $d < d_{\min}$ **then**
    $(d_{\min}, s^*) \leftarrow (d, s)$
  end if
end for
if $d_{\min} > \delta$ **then**
  return discard n
else
  return assign s^* to n
end if

Listing 1: Insert Non-Verbal Tag into Transcription

```
# Inputs:
# n: [n_s, n_e], non-verbal segment
# l: non-verbal label
# W: word list [(s_i, e_i, w_i)] for i = 1 to N
s_tag = f"[{l}]<B>"
e_tag = f"</B>"
# Find word indices overlapping with the non-verbal segment
i_s = max(i for i, (s, _, _) in enumerate(W) if s <= n_s)
i_e = min(i for i, (_, e, _) in enumerate(W) if e >= n_e)
if i_e - i_s > 1:
    # Spans multiple words
    return W[:i_s + 1] + [s_tag] + W[i_s+1:i_e] + [e_tag] + W[i_e:]
else:
    # Single word
    return W[:i_e+1] + [f"[{l}]"] + W[i_e+1:]
```

For the metrics, in addition to the standard word error rate (WER), we introduce three additional metrics focusing on the correctness of predicted tags in terms of both **type** and **position**:

- **F1-Score:** Measures the correctness of predicted tags compared to the ground truth.
- **Tag Position Distance (TPD):** To assess the positional accuracy of predicted non-verbal tags, we first align the predicted transcription $T^{(p)}$ with the reference transcription $T^{(r)}$ using the Levenshtein edit distance algorithm. This produces aligned sequences $T^{(p)}_{\text{align}}$ and $T^{(r)}_{\text{align}}$, an example is shown in Table 8. Let $P = \{p_1, p_2, \ldots, p_n\}$ and $G = \{g_1, g_2, \ldots, g_n\}$ denote the positions of the n aligned tags in the predicted and reference sequences, respectively. The TPD is defined as:

$$\text{TPD} = \frac{1}{n} \sum_{i=1}^{n} |p_i - g_i|. \tag{4}$$

  A lower TPD indicates better alignment of non-verbal tags. In our experiments, both the training and evaluation datasets contain only one tag per utterance to simplify the evaluation and more accurately reflect model performance.

- **Normalized Tag Distance (NTD):** To reduce the influence of sequence length, TPD is normalized by the length of the aligned reference transcription:

$$\text{NTD} = \frac{\text{TPD}}{L}, \quad \text{where } L = |T^{(r)}_{\text{align}}|. \tag{5}$$

B NonVerbalSpeech-38K Pipeline

B.1 Details on Incorporating Non-Verbal Tags into Speech Content

In this step, each non-verbal segment is matched with the nearest speech region obtained during preprocessing. The pseudocode for identifying the nearest speech region is provided in Algorithm 1.

To support non-verbal speech generation and understanding, it is necessary to integrate the textual content with non-verbal tags. The related algorithm is shown in the pseudocode 1. It is worth noting that state-of-the-art ASR models—including Whisper-large-v3 (Radford et al. 2023) and Paraformer-zh (Gao et al. 2022)—offer word-level timestamps. In practice, however, we have observed that these timestamps exhibit varying degrees of misalignment, with Paraformer-zh showing particularly noticeable discrepancies. We try using forced alignment tools, such as Montreal Forced Aligner (McAuliffe et al. 2017), to improve alignment. But due to the high WER in the transcriptions, no significant improvement is achieved, in some cases, the results are even worse.

C NonVerbalSpeech-38K Dataset

Using the proposed NonVerbalSpeech-38K pipeline, we construct the NonVerbalSpeech-38K dataset from a large collection of speech data crawled from diverse online sources. This diversity ensures coverage of a wide range of natural speaking styles and, importantly, includes rich and varied non-verbal phonemes. The resulting durations mostly fall within the range of 3 to 30 seconds (Figure 5), primarily because the Emilia (He et al. 2024) pipeline segments speech clips into chunks of 3–30 seconds. This segmentation operation also explains the peak near 30 seconds. Some segments slightly exceeding 30 seconds arise during Step 3 of the NonVerbalSpeech-38K pipeline (Algorithm 1), where non-verbal segments that do not fall within any speech region but are within a 1-second distance threshold are merged with the nearest speech region to maintain contextual continuity.